Machine Translation and Welsh: Analysing free Statistical Machine Translation for the professional translation of an under-researched language pair

Ben Screen, Cardiff University, UK

ABSTRACT
This article reports on a key-logging study carried out to test the benefits of post-editing Machine Translation (MT) for the professional translator within a hypothetico-deductive framework, contrasting the outcomes of a number of variables which are inextricably linked to the professional translation process. Given the current trend of allowing the professional translator to connect to Google Translate services within the main Translation Memory (TM) systems via an API, a between-groups design is utilised in which cognitive, technical and temporal effort are gauged between translation and post-editing the statistical MT engine Google Translate. The language pair investigated is English and Welsh. Results show no statistical difference between post-editing and translation in terms of processing time. Using a novel measure of cognitive effort focused on pauses, the cognitive effort exerted by post-editors and translators was, however, found to be statistically different. Results also show that a complex relationship exists between post-editing, translation and technical effort, in that aspects of text production processes were seen to be eased by post-editing. Finally, a bilingual review by two different translators found little difference in quality between the translated and post-edited texts, and that both sets of texts were acceptable according to accuracy and fidelity.

KEYWORDS
MT evaluation, cognitive effort, text production, Welsh translation, language planning.

1. Introduction

The 1950s saw the birth of Machine Translation (MT) (Hutchins 2001; Somers 2003), and, indeed, the automatic translation of natural language was one of the first tasks to which computers were applied (Pugh 1992; Hutchins 2001, 2004; Lopez 2008). Some involved in MT Research and Development in this earlier period of its history considered 'Fully Automatic High Quality MT' a complete possibility, although others took a more tempered approach and acknowledged even then the centrality of having to correct the output (Garcia 2012). That MT can benefit the professional translator, however, through this correcting of the raw output is now commonly accepted. This process of correcting the output so as to ensure it complies with the twin requirement of fidelity to the source language (SL) and grammaticality of the target language (TL) is known as 'Post-editing.' As post-editing is central to the empirical comparison carried out here, this process is first discussed. According to the translation think tank TAUS (Translation Automation Users Society), "Post-editing is the process of improving a machine-generated translation with a minimum of manual labour" (TAUS 2010). Following this definition, two main components of this important attempt to define post-editing can be gleaned, namely that the MT must be corrected to ensure grammaticality of the TL and fidelity to the SL, and secondly that this must be done in such a way that human labour is used sparingly; i.e. no unnecessary changes are made. How, then, could

this process using free Statistical Machine Translation (SMT) be of practical benefit to the professional translator, and how could these benefits be measured? Despite the fact that customised MT solutions based on purpose built corpora are becoming more popular, these are still important questions to ask given that SDL Trados Studio, Déjà Vu, Memsource, MemoQ and Word Fast Pro systems all allow users to connect to Google Translate via an API key[1]. Typically MT is used within these workbenches where the translation memory (TM) fails to offer a match of 70% or above, though the exact threshold selected is configurable. The next section reviews the available literature related to the comparison of MT Post-editing with human translation, with a view to providing a theoretical background for the current analysis. Studies that collated translation and post-editing data from students or other non-professionals were not included.

2. Comparing MT and human translation

The evaluation of MT takes a myriad of forms, and each project will have tailored its evaluation criteria according to the expected use of the system (White 2003). Accepting that the MT output rarely needs to be perfect for it to be useful Newton (1994: 4), for example, stated that:

> Direct comparisons between a system's raw output and human translation are therefore pointless; as MT is a production tool, its capacity to increase or speed up production, within acceptable cost parameters, is the only valid measure of its effectiveness.

According to Newton then, the measuring stick for the usefulness or otherwise of MT is its ability to speed up the work of translators and its capacity to increase their productivity, a metric which would likely be measured by words per minute/hour and daily throughput. Measuring increases in processing time and productivity, however, is only one approach to the evaluation of MT within the context of professional translation, and others have considered variables which arguably determine this processing time and resultant productivity. These variables include keystrokes and text production in general, as well as the cognitive challenges and benefits that the post-editing of the MT output as opposed to translation can bring. A pioneering study by Krings (2001) brought these variables together into one triadic framework, which consists of cognitive, technical and temporal effort. According to Krings (2001: 179), cognitive effort can be defined as 'the extent of cognitive processes that must be activated in order to remedy a given deficiency in a machine translation'. Technical effort in turn refers to the process of producing text and the manipulation of it on screen, and finally temporal effort refers to the time taken to complete the translation. Using time and productivity as well as the triadic framework offered by Krings (2001), a number of studies have compared variables related to these metrics between translating and post-editing, or between translating, post-editing and revising TM matches.

O'Brien (2007), Flourney & Duran (2009), Groves & Schmidtke (2009), De Sousa, Aziz & Specia (2011), Lee & Liao (2011), Green, Heer & Manning (2013), Läubli *et al*. and Aranberri *et al*. (2014) found that post-editing MT was quicker in terms of processing time than translation, and O'Brien (2007), Läubli *et al.* (2013), Carl, Gutermuth & Hansen-Schirra (2015) and Koglin (2015) found, as well as translation being speeded up, that text production in terms of keystrokes was also reduced when the MT output of the same source text was post-edited compared to the process of translating it. In terms of productivity specifically, as opposed to processing time, a range of published studies have shown within empirical frameworks that MT post-editing can boost the productivity of professional translators. Guerberof (2009, 2012), Kanavos & Kartsaklis (2010), Plitt & Masselot (2010), Federico, Cattelan & Trombetti (2012), Moran, Lewis & Saam (2014), Silva (2014) and Zhechev (2014) all report that using MT allowed the participating translators to improve their productivity over a period of time. Despite being known for its 'elusiveness' (Vieira 2014: 189), cognitive effort has been measured in a number of different ways by researchers working in Translation Studies as well as MT research. Pauses in text production, based on the work of Butterworth (1982) and Schilperoord (1996), have been used in MT research to investigate the cognitive effort invested in MT post-editing (Krings 2001; O'Brien 2006a,b; Lacruz, Shreve & Angelone 2012; Lacruz & Shreve 2014; Koglin 2015), cognitive effort in translation and the revising of TM matches (Mellinger 2014; Screen 2016), as well as to investigate other aspects of cognitive processing in translation (Jakobsen 2002, 2003, 2005; Dragsted 2005, 2006, 2012; Immonen 2006a,b; Vandepitte 2015). The original work of Butterworth (1982) and Schilperoord (1996) posited that the number and duration of pauses measured in language production can be related to processing effort of varying degrees and that Working Memory Capacity is inextricably linked to this processing effort (see Section 5.3 below where pause analyses are discussed in more detail). Another popular research method using variables related to gaze data gleaned from eye-tracking technology has been used in translation and post-editing research alike recently, analysing pupillometrics, average fixation duration or number of fixations. Pupil dilation, i.e. changes in pupil size, have been related to increases in cognitive load (Marshall, Pleydell-Pearce & Dickson 2003; Iqbal *et al.* 2005), and as such this variable has been used by researchers interested in measuring this psychological construct as it applies to translation and post-editing. O'Brien (2006c, 2008) for example measured changes in pupil dilation using eye tracking equipment as translators interacted with different percentages of TM matches, and this pupil dilation was found to be lower when the participants were asked to revise segments of a lower match value. Average fixation time and duration was also found to be lower for post-edited segments with a high GTM score (*General Text Matcher cf. Turian et al. (2003)*) and low TER score (*Translation Edit Rate cf. Snover et al. (2011)),* thus confirming the inverse relationship between increased processing speed and reduced cognitive effort (O'Brien 2011). Doherty,

O'Brien & Carl (2010) used gaze time, average fixation duration, fixation count and pupil dilation to evaluate the usability of raw MT output, and average gaze time and fixation count were found to be higher for raw MT that was rated beforehand as 'bad' than MT segments rated as 'good.' Carl, Gutermuth & Hansen-Schirra (2015) measured comparative cognitive effort between human translation and post-editing using gaze data, and the texts that were processed by translators who did not have access to a machine translated text were found to have higher fixation durations and fixation counts than those who did. Koglin (2015), who used pauses as well as average fixation duration data, found however that there were no statistically significant differences in terms of pause and fixation data between those who post-edited metaphors as compared to those who translated them in terms of both metrics. The post-editors however were found to be quicker on average than the translators who translated manually.

3. Why evaluate MT for Welsh now?

It should be clear, then, that the use of MT within the translation workflow, according to a number of published studies, can in fact decrease time spent in translation and increase processing speed and productivity, decrease those variables related to text production and is capable of decreasing cognitive effort as measured by gaze data and pause analyses. A number of studies have also shown that the final quality of the translations does not suffer despite these decreases in effort (cf. Section 7.2 for a discussion of quality). These assumptions were translated into five testable deductive hypotheses (Table 1) that were investigated using a between-groups design, having recruited professional, practising translators of the language pair investigated. It was decided that the analysis of any contribution MT could make to the translation of Welsh was timely for two reasons, and it is likely that these reasons will be familiar to minority language communities outside of Wales. First of all, scholars working in language planning have noted the important role translation plays in normalisation efforts (Gonzalez 2005; Meylaerts 2011) and have reminded us that official language policies, whether they explicitly acknowledge the fact or not, almost always lead to the practice of translation (Núñez 2013; Floran 2015). Indeed this is also the case in Wales. Efforts since the 1960s, when the British State gradually gave way to calls for greater Welsh cultural and political autonomy, and especially since 1993 when the Welsh Language Act was passed which created Wales' first professional language planning agency and led to a spike in Welsh-English translation (Kaufman 2010, 2012), professional translation has become part of the 'ethos' of Welsh societal bilingualism (Miguélez-Carballeira, Price & Kaufman 2016: 125)[2]. The current official language plan for Wales can be found in the Welsh Government's "A Living Language, A Language for Living" (Welsh Government 2012). In it, the important role translation technology can play, MT included, is given attention (Welsh Government, p. 50). This commitment to translation

technology was again confirmed in a later policy document, published in response to the UK Census figures for Welsh published in 2012 (Welsh Government 2014, p. 11).

Given the context within which translation occurs, then, and its importance to normalisation efforts for minority language communities, as well as considering the official stance of the Welsh Government in relation to automatic translation technology, an experiment was carried out to test these apparent benefits. Google Translate was chosen for two reasons. First of all, it is available as an API in the three most common TM systems in use by Welsh translators, which according to Watkins (2012) are Déjà vu, SDL Trados and WordFast. Secondly, a recent human evaluation of Google's raw MT output by five professional freelance translators of Welsh found that in terms of both accuracy and fluency, a majority of segments were found to be either 'Excellent' or 'Satisfactory' in terms of fluency, and either 'Correct' or 'Partly Correct' in terms of accuracy (Techscribe 2010). No such evaluation for any other MT system for English to Welsh could be found. The reviewers were asked to analyze a corpus of sixty sentences each, with accuracy rated as either 'Correct', 'Partly Correct', 'Incorrect' or 'Nonsense,' and with fluency rated as either 'Excellent,' 'Satisfactory,' 'Bad' or 'Incomprehensible.'

4. Evaluation criteria: hypotheses

No evaluation of Google Translate for Welsh *translators* has yet been published, and so variables that are relevant to professional translation for this language pair are yet to be considered, despite the government policies mentioned above. As Daelemans & Hoste (2009) and Ramon (2010) note, using translation as a baseline and comparing the practical benefit of post-editing against it is an essential part of MT evaluation within a professional context. The hypotheses listed in Table 1 were tested. The dependent variables measured are noted, along with a description of how the variable was measured. In terms of cognitive effort, the theory behind the metric chosen is described in the next section.

Relation to Krings (2001)	Number	Hypothesis	Dependent Variable	Variable Operationalisation
Cognitive Effort	1	Post-editors will find the task of producing a translation cognitively less demanding than the translators	All pauses between first and final keystrokes for both groups	All pauses located between first and final keystrokes for each segment by every translator in the two groups, as recorded by Translog-II and displayed in the Linear Representation. The rationale behind the exclusion of pauses before the first keystroke (associated with reading) and after the last keystroke (associated with evaluation) is outlined in the next section
Technical Effort	2	The number of alphanumeric keys struck by the post-editors will be less than those struck by translators	All letter, number and punctuation keys	All letter, number and punctuation keys recorded in the Linear Representation produced by the Translog-II software
	3	More keys related to editing extant text on screen will be struck by the post-editors than by the translators	All SPACE, Delete, arrow key, mouse clicks, Backspace and Ctrl combination keys struck	All listed keys recorded in the Linear Representation produced by the Translog-II software
Temporal Effort	4	The draft processing times of the post-editors will be significantly less than those of the translators	Time taken to complete a draft translation before final self-revision	Task time as recorded by the Replay Function in Translog-II
	5	The total processing times for the post-editors will be lower than that of the translators	Time taken to complete a finished translation	Time recorded by Translog-II from 'Start Logging' until 'Stop Logging'

Table 1. Deductive hypotheses tested

5. Methodology

The experiment was carried out in *Translog-II* (Carl 2012) and conducted at Cardiff University, UK, with the aim of evaluating Google Translate in terms of its ability to assist translation from English to Welsh. *Translog-II* is a key-logging programme which logs all keystrokes pressed during a session as well as pauses recorded between keystrokes. A secondary aim was to contribute to the evaluation literature from the perspective of an under-researched language pair, and to contribute evidence from a controlled experiment. Data was gleaned from the Linear Representation provided by *Translog-II*, as well as its Replay Function. All statistical analysis was done using IBM SPSS, the confidence threshold used was 95% and all statistical tests were two-tailed.

5.1 Participants

Ten professional translators were recruited, nine of whom were members of Cymdeithas Cyfieithwyr Cymru (*the Welsh Association of Translators and Interpreters*), membership of which is gained through passing an examination after translating professionally for at least a year. All participants, however, had a minimum of 5 years' experience in professional translation. All participants were familiar with Translation Memory tools, and all confirmed they used either Memsource (n=2) or Wordfast Pro (n=8) in their respective organisations (Cardiff University, the Welsh Government and the National Assembly for Wales). All were familiar with SMT (all were aware of Google Translate and Microsoft Translator), but no participants were trained in post-editing. Two translators in the Experimental Group had at least one year's experience of post-editing MT output, however. All participants were familiar with the text type (a general information text from a local authority), as all were familiar with and experienced in translating for the public sector domain in which this type of text is common. No participant had seen the source text beforehand.

5.2 Experimental Design

These ten translators were randomly assigned to a Control Group (CG) (n=4) who translated, and an Experimental Group (EG) (n=6) who post-edited a machine translated version of the same source text given to the CG. The source text contained 1,499 characters and 316 words, and the machine translation contained 1,566 characters and 346 words. The source text can be found in Appendix A and the raw MT output used in Appendix B. Two participants' data had to be discarded as one participant failed to save data correctly before doing the task again in half the time, and the other failed to complete the task.

5.3 Quality expectations

Post-editors were given a copy of the TAUS post-editing guidelines and were asked to correct the MT output so as to make it a translation of publishable quality, but not to make any unnecessary changes in the process of doing so. As such, a *full* post-edit was required. These guidelines were explained to participants before commencing. All translators were informed in their research ethics permission form that any set of translations may be taken for an analysis of quality by qualified professionals at a later date, and all agreed to this. All participants were aware, therefore, of the quality expected.

5.4 Apparatus

The software used to collect data was *Translog-II* as noted, and, as this research software is unfamiliar to most translators, all participants were

asked to type a short paragraph in English in the software in order to gain familiarity with how *Translog-II* looks, how to open projects and how to save files. It also allowed participants to become accustomed to a new keyboard and a different machine. However, all participants use desktops in their own work and so all were familiar with this type of workstation. In terms of the CG, the English source text was shown on the left and the target text window on the right within the *Translog-II* interface, using its parallel screen option. This was done for both groups so as to increase ecological validity as TM systems typically display the source text on the left and the target text on the right similar to a bitext. Participants were asked to click 'ESC' in order to be able to see the next segment. Participants were asked not to proceed to the next segment until they had finished the previous one. In terms of the EG, all 15 source segments were displayed on a parallel screen, but in order to see the next MT segment the participants were required to strike 'ESC' also. The participants in the EG were asked not to press 'ESC' until they had finished processing the previous MT segment. The source text side was locked for both groups. The parallel layout chosen for the *Translog-II* GUI therefore was kept constant for translators and post-editors.

5.5 Pauses and cognitive effort

Research that has relied on pauses as a metric to gauge cognitive effort was outlined above. Whilst accepting that supplementary methods should ideally be taken advantage of when using pause analysis to measure cognitive effort (O'Brien 2006: 1), pauses are used here as the use of eye-tracking equipment to collect gaze data was not possible at the time of data collection and subjective ratings of effort have been shown to be inaccurate by past research (Koponen 2012; Gaspari *et al*. 2014; Teixeira 2014; Carl, Gutermuth & Hansen-Schirra 2015; Moorkens *et al.* 2015). Previous collated data regarding translator's attitudes towards MT technology has also shown that some can be negative towards it (Guerberof 2013), and as such this was another reason not to rely on qualitative data as antagonisms could well have biased the participant's ratings. It was also noted above that pauses in language *production*, according to the theories of Butterworth (1980) and Schilperoord (1996), are linked to cognitive effort. Butterworth (1980: 156), spelled it out: "The more the delays [e.g. pause time], the more cognitive operations [e.g. processing effort] are required by the output." Schilperoord (1996: 11) adopted a similar stance, "[…] longer pauses reflect cognitive processes that are *relatively* [emphasis by Schilperoord] more effortful compared to processes reflected by shorter pauses." Kumpulainen (2015: 55), commenting from within Translation Studies in particular, also succinctly explained the logical basis of pause analysis and the distinction between pause number and pause duration, thereby linking the theories of Butterworth and Schilperoord, "Several 1-second pauses clustered in a segment can be regarded as an equally clear indication of extra cognitive effort as one long pause in a segment." Pauses,

according these researchers, are a function of Working Memory Capacity (WMC). As working memory is considered to be *limited* (Baddeley 1999, 2007), and in particular the Central Executive that is responsible for managing and allocating cognitive resources and the Phonological Loop which is linked to language production including writing (Kellog 2001; Kellog *et al*. 2007), it follows that translation and post-editing, as linguistic tasks, are also subject to these constraints on working memory[3]. Lacruz & Shreve (2014: 248) also invoke WMC as a central element of cognitive effort in post-editing:

> The Cognitive Effort exerted by post-editors will be driven by an interaction between the internal factors of their available cognitive resources and the extent to which they need to be allocated…and the external factor of MT quality.

Segments that contain SL material and MT output that require deeper, more effortful processing on the part of the translator will likely cause greater pause activity. This is arguably because the translator and post-editor will be forced to concentrate on singular challenging elements of the segment at a sub-sentential level, and is required to do this whilst keeping in mind the whole due to WMC (Dragsted 2005: 50). Disregarding for the moment any time spent pausing at the beginning of segments, as well as time spent revising extant text already produced, it could be postulated that internal pauses recorded between initial reading and segment revision are likely to be correlated with cognitive problem solving within that particular segment. Were the duration of internal pauses in a segment to be high, then one could argue that due to WMC the translator divided attention at different times between different translation and post-editing related problems, and so the cognitive effort related to this particular segment could be considered high. Shreve and Diamond (1997: 243) seem to provide support for this understanding:

> Frontal systems may employ the Central Executive to initiate more effortful processing and allocate attentional and other resources when we become aware of what we might loosely call ‘’information processing problems.’

This will, in turn, lead to "reductions in the efficiency with which a particular task is performed" (Shreve and Diamond *Ibid*.), i.e. it will be slowed down as a result of frequent pausing and by longer fixations on particular elements within the text. As a result, pauses were decided upon as a measure of cognitive effort and a threshold of 1.5 seconds was used, pauses below which were not analysed. It was decided to use this threshold as it has been linked to the theoretical capacity of the Phonological Loop (Dragsted 2006), a component of working memory which, as noted above, is linked to written language production.

Pauses before the first keystroke and after the last keystroke were not included in the analysis of pauses linked to cognitive effort. Translators are unlikely to *produce* text without having read any of the source text at all,

and post-editors are just as unlikely to start the correction process before having decided what the problem is. Equally, all pause activity after text production stops once a full rendition has been created is unlikely to be linked to translation or post-editing, and much more likely to be related to self-revision. Recent research has provided evidence that more effort is expended by translators in transfer processes than in the orientation and revision phases (Jakobsen & Jensen 2008; Sharmin, Spakov & Räihä 2008; Pavlović & Jensen 2009), and so arguably any attempt to isolate the pauses linked to these different processes would be desirable. As noted by Jensen (2011, p. 49), "Assuming that processing effort is identical across all three production stages entails the risk of basing an analysis on data that reflect several tasks." This tripartite task sequence of reading (and possibly *starting* to form a mental translation (cf. Macizo & Bajo 2004; Ruiz *et al.* 2008; Jakobsen & Jensen 2008; Dragsted 2010), followed by text production and finally revision has also been found for post-editing, as the reading and evaluation data noted above suggest. Carl *et al*. (2011) found in their pilot study that the gaze data of post-editors showed that post-editors also tended to read the source text as did translators (albeit to check the target text against the source text as well as reading for meaning), and this was confirmed in a later study by Carl, Gutermuth & Hansen-Schirra (2015). Including this reading and evaluation time could therefore have skewed the conclusions drawn in relation to cognitive effort and so these data were treated separately. In fact, total reading time for the post-editors was found to be longer than that of the translators (CG M= 6, Mdn= 4.5, EG M= 15, Mdn= 12) which was a statistically significant difference according to a two-tailed bootstrapped independent T-test (-9.13, BCa 95% CI (-12.51, -5.76), t(118)= -5.36, p= .001). This was also true for evaluation time (CG M= 3.4, Mdn= 0, EG M= 8, Mdn= 4.5), which was also significant according to the same test (-4.77, BCa 95% CI (-8.75, -.77), t(118)= -2.4, p= 0.25).

Finally, in terms of other pause metrics used in the literature, Pause Ratio (O'Brien 2006), Average Pause Ratio (Lacruz & Shreve 2014; Mellinger 2014) and Pause to Word Ratio (Lacruz & Shreve 2014) were not used. This was decided as they analyse all pause data collected together as though it refers to one single process of equally distributed effort, even though the data most likely represent three distinct cognitive processes all of which require different processing demands.

6. Results

With the methodology outlined and the cognitive theory that was operationalised in the pause analysis also described, the next section discusses the results obtained.

6.1 Cognitive effort

The hypothesis in relation to this evaluation parameter was that the post-editing of MT output could render the process of finding TL material cognitively easier for the EG. In this case the Alternative Hypothesis had to be accepted, as a statistical difference was found between the internal pauses produced by the CG compared to the EG (CG M=16, Mdn= 10, EG M= 32, Mdn= 20). The EG paused significantly longer between reading for meaning (orientation) and self-evaluating than the CG, and so it cannot be argued in this case that post-editing was cognitively easier. Given the non-parametric nature of the data, a boot-strapped independent T-test was used, which gave -.164, BCa 95% CI (-28.84, -5.96), t(77)= -2.71, *p*= .015. This is displayed graphically in Figure 1 below, along with the pause data for Orientation and Self-revision.

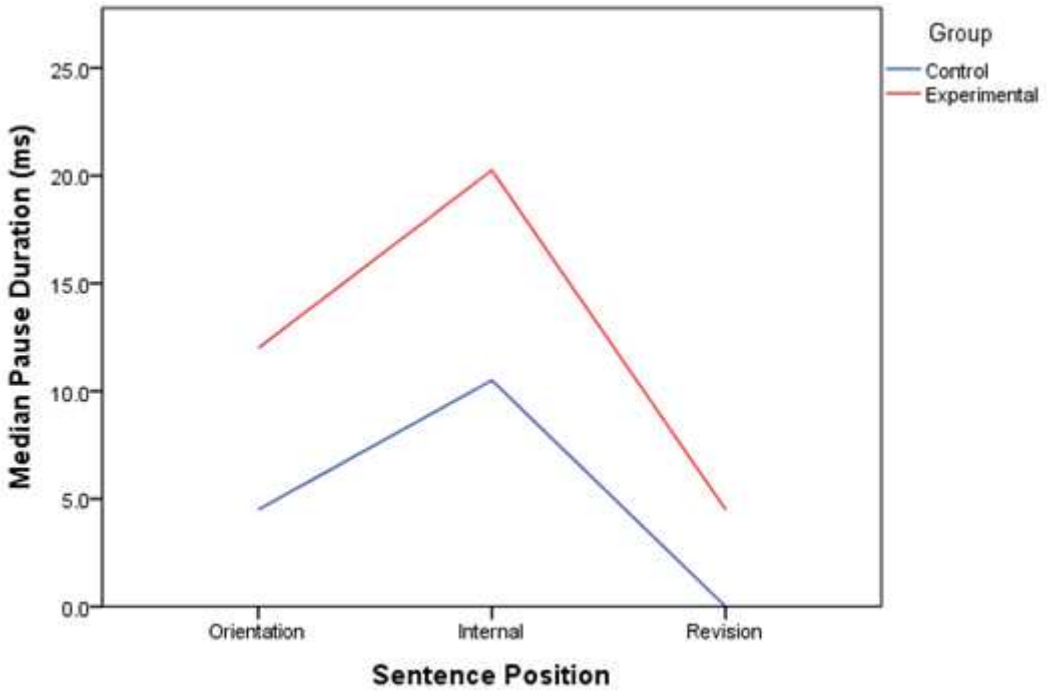


Figure 1. Average pause duration (ms) for both groups in all pause positions

Contrary to the studies cited above that found that post-editing can in fact be cognitively easier than translation, this finding seems to contradict this assertion.

6.2 Text production

Hypothesis 2 predicted that the translators would in fact produce more alphanumeric characters as text would need to be typed anew. The Null Hypothesis was rejected in this case. The CG produced more alphanumeric characters during text production processes (M= 114, Mdn= 114) than their EG counterparts (M= 18, Mdn= 12.5), and this difference was statistically significant. Bootstrapping was used when performing the independent T-

test due to the non-parametric nature of the data (96, BCa CI 95% (81.7, 112.1), t(12.6)= 70, *p*= .000). It can also be deduced from this finding that the post-editors in the EG tended not to rewrite translations. This difference is represented in Figure 2 below, where this significant difference can clearly be seen:

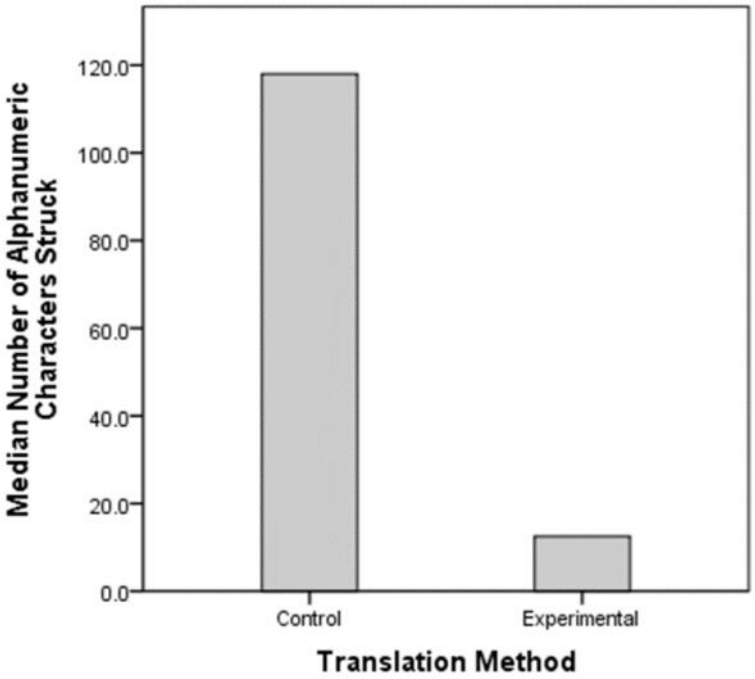


Figure 2. Median number of alphanumeric characters produced

It was also predicted, in Hypothesis 3, that post-editors would rely more on non-alphanumeric keys however as a result of their *adaption* and *manipulation* of text, rather than their typing it anew. This form of text production will lead to mouse-clicks, deletions by using the Backspace key and Delete key, copy and paste procedures, use of the arrows and spacebar as well as Ctrl commands. The adaption, correction and manipulation of extant text on screen will also be a feature of manual translation processes. Translators, having written a first rendition, may decide to adapt and improve upon their first attempt or delete it altogether. This latter tendency is known as a 'false start.' Post-editors however will likely rely more on this form of text production as they have to *correct* text rather than type their own on a blank screen, unless they delete the MT output and translate independently. There was a difference between groups (CG M= 27, Mdn= 9, EG M= 43, Mdn= 18), although the Null Hypothesis had to be accepted in this case as no statistically significant difference between the datasets was found according to a bootstrapped independent T-test: -16.4, BCa CI 95% (-37.5, 2.24), t(77)= -1.73), *p*= 0.088). The low p-value here however (0.088) means that statistical significance was marginally missed.

These results show that the underlying text production processes in translation and post-editing are likely to be different. The results presented here also confirm a similar finding by Carl, Gutermuth & Hansen-Schirra

(2015), who found that the number of insertions and deletions (measured by a novel metric labelled 'InEff,' or Inefficiency Metric) was higher for translators than it was for post-editors of the same MT system studied here (Google Translate), as well as earlier results of another experiment carried out by Carl *et al.* (2011). These scholars found that post-editing led to more deletions, navigation keystrokes and mouse clicks than translation, but less insertions. It should be noted however that the number of alphanumeric characters produced by the post-editors, arguably the most common form of text production, was significantly less and the total number of keystrokes produced by the CG was 8,064 but 3,026 for the EG. These are interesting findings despite the use of different language pairs.

In Table 2 below, summary statistics are offered for the total of all keystrokes recorded in every segment. The 'Other Keys' column refers to use of arrows, the space bar and CRTL combinations.

	Control Group (Translation)					**Experimental Group (Post-editing)**				
Seg.	Alphanu-meric Characters	Dele-tions	Mouse Clicks	Oth-er Keys	Total Strokes	Alphanu-meric Characters	Dele-tions	Mouse Clicks	Oth-er Keys	Total Strokes
1	89	32	6	0	127	10	4	7	0	21
2	421	25	4	0	450	5	4	3	0	12
3	449	56	6	2	513	51	38	18	4	111
4	713	89	10	6	818	128	83	24	46	281
5	332	26	4	29	391	14	9	7	0	30
6	556	50	3	22	631	157	142	43	94	436
7	181	23	3	2	209	94	17	7	52	170
8	409	65	8	149	631	104	69	14	142	332
9	561	100	3	2	666	196	122	26	151	495
10	745	88	9	4	846	52	39	11	72	174
11	484	29	3	22	538	79	94	11	107	291
12	574	98	4	2	678	17	7	13	6	43
13	810	110	11	51	982	120	75	24	238	457
14	67	9	3	1	80	0	0	0	0	0
15	461	35	3	5	504	44	34	17	78	173
Total	6852	835	80	297	8064	1071	737	225	990	3026
Mean	457	55.7	5.3	19.8	537.6	71	49	15	66	202
Median	461	50	4	4	538	52	38	13	52	173
S. Dev	214.6	32.4	2.7	37.2	250.6	58	44	11	69	166

Table 2. Total of all keystrokes for each segment

6.3 Processing time

Finally, differences in processing time will be discussed. Productivity is not analysed due to the relatively modest number of segments processed (120). Taking first of all the processing times recorded before the process of final revision (called here 'draft processing times'), where every

participant went through a completed draft making a small number of changes to some segments, no statistical differences could be discerned between the two groups[4]. These draft processing times were captured using the replay function in *Translog-II*. The central tendencies of the time taken to complete a segment on average by both groups were similar: CG Mean = 63 seconds, SD = 39, CI (53, 74), EG Mean = 62 seconds, SD = 41, CI (51.8, 73.2). As the processing times for each segment processed by each participant was the only dataset to have a normal distribution, an independent T-test was used without bootstrapping to test for differences between means. The result ($t(181) = 118$, $p = .857$) showed no statistically significant difference between the draft processing times for both groups, as to be expected given the almost identical means. The Effect Size was also small ($d = 0.1$). The Null Hypothesis was accepted as a result and deductive Hypothesis 4 was rejected. These data are illustrated in Figure 3 below:

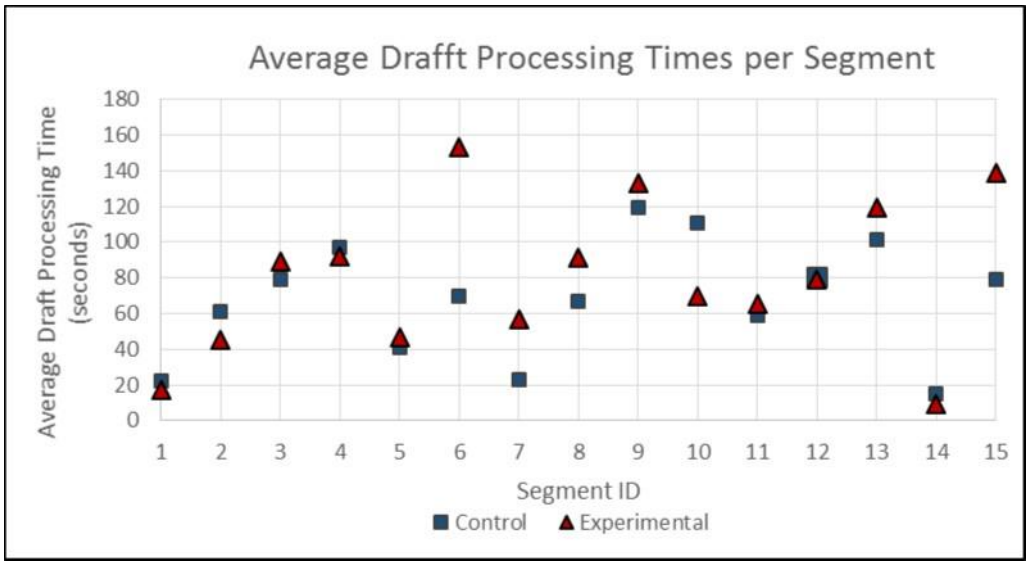


Figure 3. Average draft processing times per segment

In terms of total processing time, i.e. the time recorded from the start of the session in *Translog-II* to the time 'Stop Logging' was pressed by the participant after completing final self-revision, post-editing did not lead to automatic reductions in the time taken to complete a professional translation. It should be noted however that P5, P6 and P9 in the EG were quicker than P1 in the CG, and that P9, the quickest post-editor, was 5.17 minutes quicker than P1 and 40 seconds quicker than P2. Also, taking the average time taken to process each segment by both groups, it should be noted that in 33.3% of cases, the machine translated segments were processed faster than when the same segments were translated by the translators who *didn't* have access to MT. Equally however, translation was quicker on average in the remaining 67.7% of cases. Hypothesis 5 then must be rejected. These differences between Total Processing Times are displayed graphically in Figure 4 below. Red bars denote CG data and blue bars denote EG data.

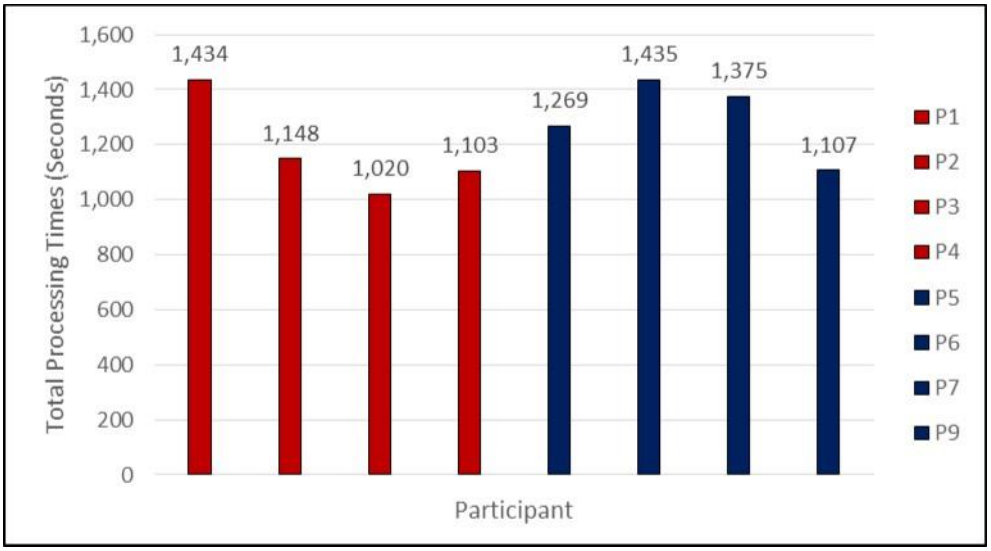


Figure 4. Total processing times (seconds)

7 Quality assessment of final texts

In order to validate the data analysed above, a quality assessment (QA) was carried out on the completed translations. The methodology used is described below.

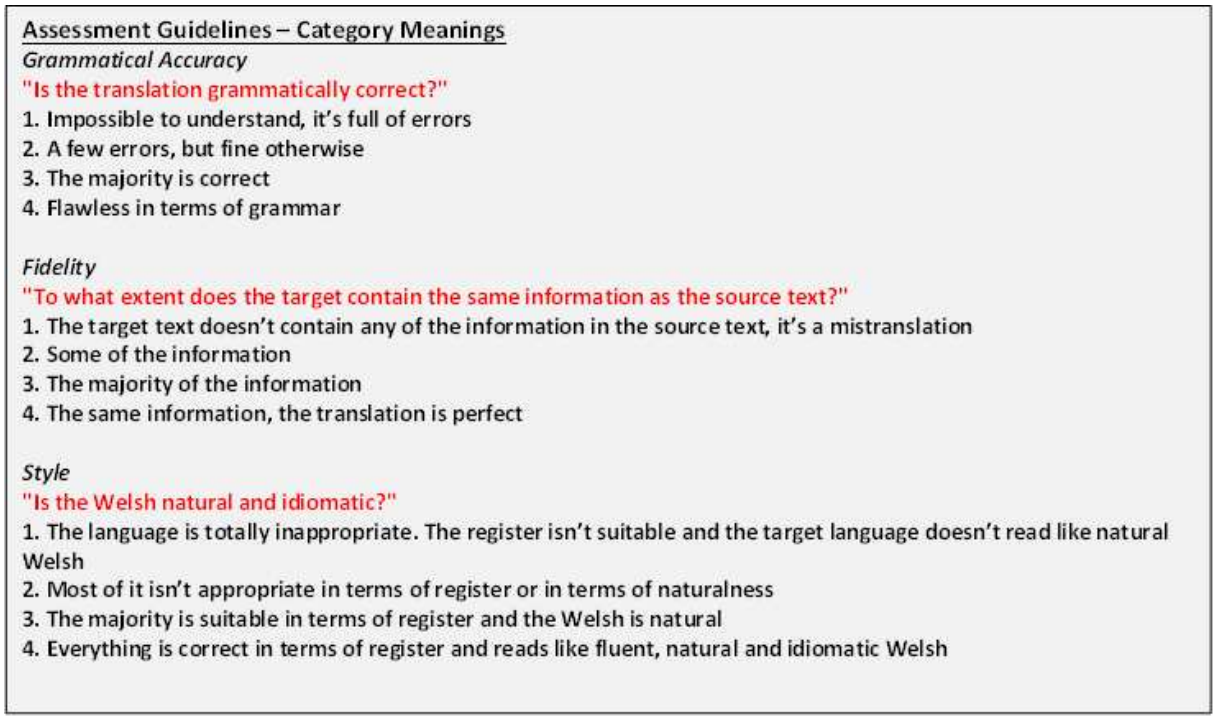
7.1 Reviewer profile

Two experienced reviewers, both of whom had at least 20 years' experience in the translation industry and both of whom were members of Cymdeithas Cyfieithwyr Cymru, carried out a bilingual evaluation of the translated and post-edited texts. One held the position of senior translator before starting their own language company, and the other was head of the translation unit in their respective organization. Both also held a PhD in translation or linguistics. This aspect of the methodology then complies with Hansen's advice (2009: 394) on product validation in translation process research, which is that qualified translators with experience of revision should ideally be used. In terms of the number of reviewers, it is noted that published studies by Garcia (2011), Teixeira (2011) and Läubli *et al*. (2013) also used two reviewers.

7.2 Method

Reviewers were sent an Excel spreadsheet containing all translations produced by participants in the CG (Set A) in one tab, and all translations produced by the EG (Set B) in another tab. This review was blind; no reviewer knew which tab contained which set of translations. Reviewers were asked to rate the translations of both groups against the source text (Appendix A), using a scale of 1-4, where 1 was the poorest score and the 4 the best. Each translation was judged for grammatical accuracy, fidelity and style. The criteria used in Fiederer & O'Brien's 2009 study were

therefore replicated despite a change in terminology; grammatical accuracy corresponds to clarity and fidelity to accuracy. A consideration of register was also added to 'Style.' The definitions used are shown in Figure 5. No reviewer knew which set was produced through manual translation and which was produced through post-editing.


Assessment Guidelines – Category Meanings

Grammatical Accuracy

"Is the translation grammatically correct?"

1. Impossible to understand, it's full of errors
2. A few errors, but fine otherwise
3. The majority is correct
4. Flawless in terms of grammar

Fidelity

"To what extent does the target contain the same information as the source text?"

1. The target text doesn't contain any of the information in the source text, it's a mistranslation
2. Some of the information
3. The majority of the information
4. The same information, the translation is perfect

Style

"Is the Welsh natural and idiomatic?"

1. The language is totally inappropriate. The register isn't suitable and the target language doesn't read like natural Welsh
2. Most of it isn't appropriate in terms of register or in terms of naturalness
3. The majority is suitable in terms of register and the Welsh is natural
4. Everything is correct in terms of register and reads like fluent, natural and idiomatic Welsh

Figure 5. Assessment criteria used for the ratings

7.2 Quality assessment results

All three parameters (grammatical accuracy, fidelity and style) were taken together and the average calculated using the mean in order to gain a single translation quality score for each translator in both groups. This result is shown in Figure 6 below. As can be seen, there is little difference in this overall translation quality score and all participants had a mean score of 4 which was the highest score possible. This lack of difference was then confirmed by a Wilcoxon signed rank test (p = .162).

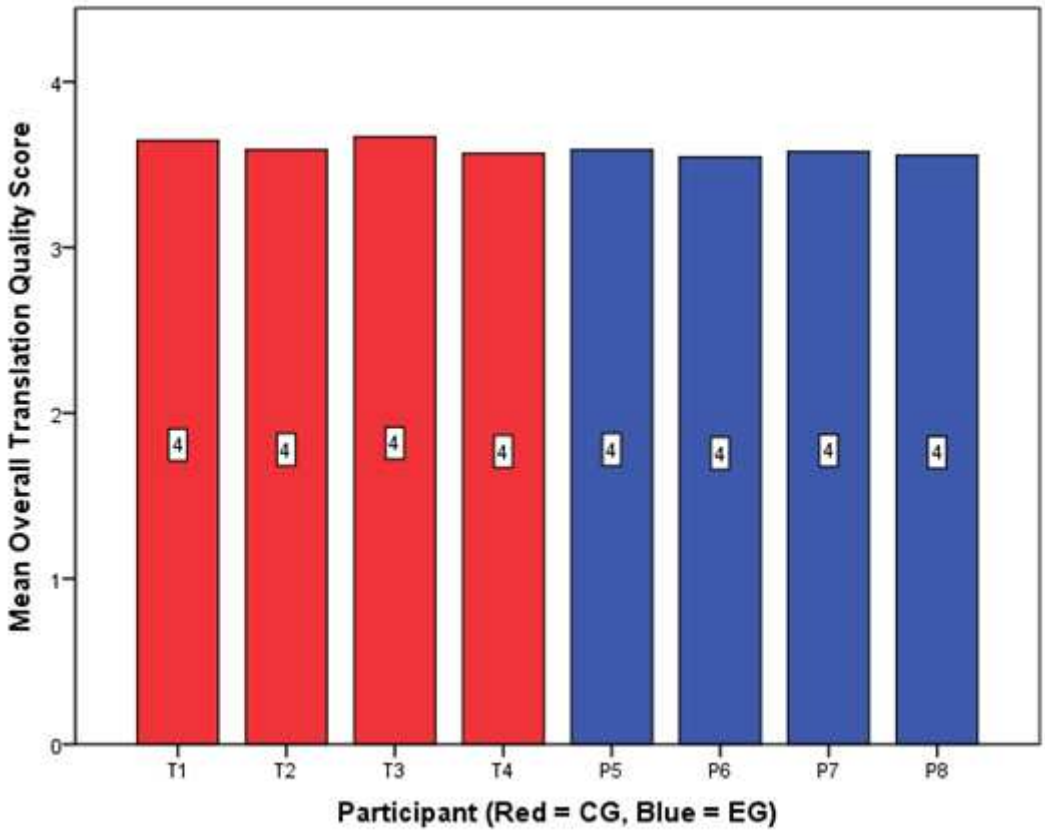


Figure 6. Overall Translation Quality Score for both groups

Looking individually at the three parameters combining the scores given by both reviewers, the same tendency is observed for grammatical accuracy (p = .059) and fidelity (p = .094), but not for style (p = .028) according to a Wilcoxon signed rank test. This result for style (p = .028) in particular is interesting, as Fiederer & O'Brien (2009) found that their participants also rated the translations produced manually higher for style, with a statistically significant difference being found using a Wilcoxon signed rank test. It should be noted however that the mean score for the EG was 3.53 and 3.68 for the CG, so despite this difference the machine translated texts still succeeded in gaining a fairly high score overall for style. What these results also do is validate the data gleaned for the results presented above on effort; as all translations were deemed acceptable by experienced peers, it can be said that all translators treated the task as they would any other.

The results of this QA then by two experienced reviewers show that the use of post-editing when producing a translation does not necessarily lead to texts of inferior quality, which echoes results of other studies that have investigated this aspect of translation and post-editing. Despite using dissimilar methodologies in their QA, Bowker (2009), Bowker & Ehgoetz (2009), Fiederer & O'Brien (2009), Garcia (2011), Plitt & Masselot (2010), Carl *et al.* (2011), Teixeira (2011), Läubli *et al*. (2013), O'Curran (2014) and Ortiz-Boix & Matamala (2016) also found that post-editing does not lead to a decrease in translation quality, and in fact Garcia (2010) and Plitt

& Masselot (2010) actually found that the post-edited texts were deemed to be of *higher* quality than the translations.

8 Conclusions

Despite a body of research which has shown that MT in many cases can speed up translation, make the translation process cognitively and practically easier as well boost productivity (accepting that the results are system, domain and language dependent), this study has shown that Google Translate did not help the four post-editors who were recruited here for the EN-CY language pair, and no statistical differences could be found overall between groups in terms of time and cognitive effort. This finding chimes with that of Carl *et al*. (2011) who could find no statistically significant differences using Google Translate for the pair EN-DA in terms of processing time (also using participants without post-editing training), and Sekino (2015) who found that the same system did not significantly speed up translation for PT-JA either. It should be borne in mind however that one aspect of the translation process was made easier by post-editing; the text production process in terms of the total number of keystrokes recorded was less in the case of the post-editors. It is not argued here that MT as a translation strategy is unlikely to be useful; the research reviewed above would render such a conclusion dumbfounded. As Ramos (2010) reminds us, "MT is here to stay." This particular system however according to the evaluation criteria laid down mostly failed to be of practical benefit, but further work is required before final conclusions can be made and the sample size means findings should be interpreted carefully. One further finding however is that there is no difference in quality between translated and post-edited texts; despite marginal differences between groups in terms of time when post-editing, the fact the quality is the same in terms of grammatical accuracy and fidelity puts the use of MT in a professional context in a much more positive light and confirms the finding of a number of similar studies where the final quality of translated and post-edited texts has been compared.

9 Limitations

A limitation of this study is that only two out of four post-editors had experience in post-editing; an avenue for future research is to empirically investigate the interface between training, positive dispositions and post-editing efficiency (as done recently by De Almeida (2013)) to ascertain whether this could have affected the results presented here. Scaling up the study with a greater number of participants is also needed, as the sample of 8 here is arguably small (although less than ten participants have been used in similar research, cf. Guerberof (2009)).

Bibliography

- **Allen, Jeffery** (2003). "Post-editing." Harold Somers (ed) (1992). *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamin's Publishing Company, 297-319.

- **Alves, Fabio & Vale, Daniel Couto** (2011). "On Drafting and Revision in Translation: A Corpus Linguistic Oriented Analysis of Translation Process Data." T3: Computation, Corpora, Cognition 1(1), 105-122.

- **Aranberri, Nora** *et al*. (2014). "Comparison of post-editing productivity between professional translators and lay users." Sharon O'Brien, Michel Simard & Lucia Specia (eds) (2014) *Proceedings of the Third Workshop on Post-Editing Technology and Practice*. Vancouver, Canada: AMTA, 20-33.

- **Baddeley, Alan** (1999). *Essentials of Human Memory*. Hove: Psychology Press.

- — (2007). *Working Memory, Thought and Action*. Oxford: Oxford University Press.

- **Bowker, Lynne** (2009). "Can Machine Translation meet the needs of official language minority communities in Canada? A recipient evaluation." *Linguistica Antverpiensia* 8, 123-155.

- **Bowker, Lynne & Ehgoetz, Melissa** (2009). "Exploring User Acceptance of Machine Translation Output." Dorothy Kenny & Ryou Kyongjoo. (eds) (2009). *Accross Boundries: International Perspectives on Translation Studies*. Newcastle: Cambridge Scholars Publishing: Newcastle, 209-225.

- **Butterworth, Brian** (1980). "Evidence from Pauses in Speech." Butterworth, Brian (ed.) (1980). *Language Production. Volume 1- Speech and Talk*. London: Academic Press, 145-174.

- **Carl, Michael** (2012). "Translog - II: a Program for Recording User Activity Data for Empirical Reading and Writing Research." Paper presented at the *Eight International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 23-25).

- **Carl, Michael *et al***. (2011). "The Process of Post-Editing: A pilot study." In: Copenhagen Studies in Language, pp. 131–142. http://mt-archive.info/NLPSC-2011-Carl-1.pdf (consulted 11.03. 2016)

- **Carl, Michael, Gutermuth, Silka & Hansen-Schirra,** Silvia (2015). "Post-editing Machine Translation: Efficiency, Strategies and Revision Processes in Professional Translation Settings." Aline Ferreira & John W. Schwieter (eds) (2015). *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamin's Publishing Company, 145-175.

- **Daelemans, Walter & Hoste, Véronique** (2009). "Introduction. Evaluation of Translation Technology." *Linguistica Antverpiensia* 8, 9-17.

- **De Almeida, Giselle** (2013). *Translating the Post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages*. PhD Thesis. Dublin City University.

- **Drugan, Joanna** (2013). *Quality in Professional Translation: Assessment and Improvement*. London: Bloomsbury.

- **Sousa, Sheila C. M., Wilker Aziz and Lucia Specia** (2011). "Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD subtitles." Galia Angelova *et al.* (eds) (2011) *Proceedings of the Recent Advances in Natural Language Processing Conference*. Hissar, Bulgaria: RANLP 2011 Organising Committee, 97–103.

- **Doherty, Stephen, O'Brien, Sharon & Carl, Michael** (2010). "Eye-tracking as an MT evaluation technique." *Machine Translation* 24(1), 1-13.

- **Dragsted, Barbara** (2005) Segmentation in Translation: Differences across levels of experience and Difficulty. *Target*, 17(1), 49-70.

- — (2006) "Computer aided translation as a distributed cognitive task." *Pragmatics and Cognition* 14(2), 443-463.

- — (2010). " Co-ordination of Reading and Writing Processes in Translation: An Eye on Unchartered Territory." Gregory Shreve & Eric Angelone (eds) (2010). *Translation and Cognition*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 41-63.

- — (2012). "Indicators of Difficulty in Translation – Correlating Product and Process Data." *Across Languages and Cultures* 13(1), 81-98.

- **Dragsted, Barbara & Carl, Michael** (2013). "Towards a Classification of Translation Styles based on Eye-Tracking and Key-logging Data." *Journal of Writing Research* 5(1), 133-158.

- **Federico Marcelo, Cattelan, Alessandro & Trombetti, Marco** (2012). "Measuring user productivity in machine translation enhanced computer assisted translation." Paper presented at the *Second International Conference of the Association for Machine Translation in the Americas*. (San Diego, October 24-27 2012).

- **Fiederer, Rebecca & O'Brien, Sharon** (2009). "Quality and machine translation: A realistic objective?" *JoSTrans, The Journal of Specialised Translation* 11, 52-72.

- **Flournoy, Raymond & Duran, Christine** (2009). "Machine Translation and Document Localization at Adobe: From Pilot to Production." Paper presented at the *12th MT Summit* (Ottawa, August 26-30, 2009).

- **Folaron, Debbie** (2015). "Introduction: Translation and minority, lesser-used and lesser-translated languages and cultures." *JoSTrans, The Journal of Specialised Translation* 24, 16-27.

- **Gaspari, Federico *et al.*** (2014). "Perception vs Reality: Measuring Machine Translation Post-Editing Productivity." Paper presented to the *Third Workshop on Post-Editing Technology and Practice (WPTP-3), within the eleventh biennial conference of the Association for Machine Translation in the Americas (AMTA-2014)*. (Vancouver, BC, Canada. October 26).

- **Garcia, Ignacio** (2011). "Translating by post-editing: Is it the way forward?" *Machine Translation* 25(3), 217-237.

- — (2012). "A brief history of postediting and of research on postediting." Anthony Pym and Alexandra Assis Rosa (eds) (2012). *New Directions in Translation Studies.* Special Issue of *Anglo Saxonica* 3(3), 292–310.

- **Green, Spence, Heer, Jeffrey & Manning, Christopher** (2013). "The Efficacy of Human Post-Editing for Language Translation*."* Paper presented at the *Conference on Human Factors in Computer Systems* (Paris, 27 April-2 May 2013).

- **Goves, Declan & Schmidtke, Dag** (2009). Indentification and analysis of post-editing patterns for MT. Paper presented at the *Proceedings of the 12th Machine Translation Summitt* (Ottawa August 26-30, 429-436).

- **González, Marta García** (2005). "Translation of minority languages in bilingual and multilingual communities." Albert Branc hadell & Lovell Margaret West (eds) (2005). *Less Translated Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 105-125.

- **Guerberof, Ana** (2009). "Productivity and quality in the post-editing of outputs from translation memories and machine translation." *International Journal of Localization 7*(1), 11-21.

- — (2012). *Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation*. PhD Thesis. Universitat Rovira i Virgili.

- — (2013). "What do professional translators think about post-editing?" *JoSTrans, The Journal of Specialised Translation* 19, 75-95

- **Hansen, Gyde** (2009). "Some thoughts about the evaluation of translation products in empirical translation process research." Inger Mees, Fabio Alves & Susanne Göpferich (eds) (2009). *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*. Copenhagen: Samfundslitteratur, 389-403.

- **Hutchins, John** (2001). "Machine translation over fifty years." *Histoire, Epistemologie, Language* 23(1), 7-31.

- — (2004). "*Machine Translation and the Welsh Language: The Way Forward."* Cardiff: Welsh Language Board.

- **Immonen, Sini** (2006a). "Unravelling the Processing Units of Translation." *Across Languages and Cultures* 12(2), 235-257.

- — (2006b). "Translation as a Writing Process: Pauses in Translation versus Monolingual Text Production." *Target* 18(2), 313-335.

- **Iqbal S. *et al*.** (2005) Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. Paper presented at the *Human Factors in Computing Systems Conference* (Portland, April 2-7 2005).

- **Jakobsen, Arnt Lykke** (2002). Translation Drafting by Professional Translators and by Translation Students. Gyde Hansen (ed.) (2002). *Empirical Translation Studies*. Copenhagen: Samfundslitteratur, 191-204.

- — (2003). "Effects of Think Aloud on Translation Speed, Revision and Segmentation." Fabio Alves (ed.) (2003). *Triangulating Translation: Perspectives in Process Orientated Research*. Amsterdam/Philadelphia: John Benjamin's Publishing Company, 69-97.

- — (2005). "Investigating Expert Translators' Processing Knowledge." Helle Dam, Jan Engberg & Heidrun Gerzymisch-Arbogast (eds) (2005). *Knowledge Systems and Translation*. Berlin/New York: Mouton de Gruyter, 173-193

- **Jakobsen, Arnt Lykke & Jensen, Kristian** (2008). "Eye movement behaviour across four different types of reading task." Susanne Göpferich, Arnt Lykke Jakobsen and Inger Mees (eds) (2008). *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*. Copenhagen: Samfundslitteratur, 103-124.

- **Jensen, Kristian** (2011). *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-logging Study*. PhD Thesis. Copenhagen Business School.

- **Kanavos, Panagiotis & Kartsaklis, Dimitrios** (2010). "Integrating Machine Translation with Translation Memory: A Practical Aprroach." Paper presented at the *Second Joint EM+/CNGL Workshop ''Bringing MT to the User: Research on Integrating MT in the Translation Industry''* (Denver Colorado 4 November 2010).

- **Kaufmann, Judith** (2010). "Cyfieithu a Pholisi Iaith." *Contemporary Wales* 23(1), 171-183.

- — (2012). "The Darkened glass of bilingualism? Translation and interpreting in Welsh language planning." *Translation Studies* 5(3), 327-344.

- **Kellog, Ronald** (2001). "Competition for Working Memory among writing processes." *The American Journal of Psychology* 114(2), 175-191.

- **Kellog Ronald, Olive Thierry & Piolat, Annie** (2007). "Verbal, Visual and Spatial Working Memory in Written Language Production." *Acta Psychologica* 124, 382-397.

- **Koglin, Arlene** (2015). "An Empirical Investigation of Cognitive Effort Required to Post-edit Machine Translated Metaphors Compared to the Translation of Metaphors." *Translation and Interpreting* 7(1), 126-141.

- **Koponen, Maarit** (2012). "Comparing human perceptions of post-editing effort with post-editing operations." Paper presented at the *7th Workshop on Statistical Machine Translation* (Montreal, June 7-8, 2012).

- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (ed. Geoffrey S. Koby). Kent, Ohio: Kent State University Press.

- **Kumpulainen, Minna** (2015). "On the Operationalisation of 'Pauses' in Translation Process Research." *Translation and Interpreting* 7(1), 47-58.

- **Läubli, Samuel *et al*.** (2013). "Assessing post-editing efficiency in a realistic translation environment". Paper presented to the *European Association of Machine Translation MT Summit XIV - Workshop on Post-editing Technology and Practice* (Nice, September 2 2013).

- **Lacruz, Isabel, Gregory M. Shreve and Erik Angelone** (2012). "Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012) *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. Association for Machine Translation in the Americas, 21–30.

- **Lacruz, Isabel & Shreve, Gregory** (2014). "Pauses and Cognitive Effort in Post-Editing." O'Brien, Sharon *et al*. (eds) (2014). *Post-Editing of Machine Translation:*

Processes and Applications. Newcastle: Cambridge Scholars Publishing, 246-274.

- **Lee, Jason & Liao, Posen** (2011). "A Comparative Study of Human Translation and Machine Translation with Post-editing." *Compilation and Translation Review* 4(2), 105-149.

- **Lopez, Adam** (2008). "Statistical Machine Translation." *ACM Computing Surveys* 40(3), 1-49.

- **Marshall, Sandra, Pleydell-Pearce, Christopher & Dickson, Blair** (2003). "Integrating Psychophysiological Measures of Cognitive Workload and Eye movements to Detect Strategy Shifts." Paper presented at the *36th Hawaii International Conference on System Sciences* (Big Island, Hawaii, 6-9 January 2003).

- **Macizo, Pedro & Bajo, Teresa** (2004). "When Translation Makes the Difference: Sentence Processing in Reading and Translation." *Psicológica* 25, 181-205.

- **Meylaerts, Reine** (2011). "Translational Justice in a Multilingual World: An Overview of Translational Regimes." *Meta* 56, 743-757.

- **Mellinger, Christopher** (2014). *Computer-Assissted Translation: An Investigation of Cognitive Effort*. PhD Thesis. Kent State University.

- **Miguélez-Carballeira, Helena, Price, Angharad & Kaufman, Judith** (2016). "Introduction: Translation in Wales: History, theory and approaches." *Translation Studies* 9(2), 125-136.

- **Moorkens, Joss *et al.*** (2015) "Correlations of perceived post-editing effort with measurements of actual effort." *Machine Translation* 29(3), 267-284.

- **Moran, John, Lewis, David & Saam, Christian** (2014). "Analysis of Post-editing Data: A Productivity Field Test using an Instrumented CAT Tool." Sharon O'Brien *et al.* (eds)(2014). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing, 126-147.

- **Mossop, Brian** (2013). *Revising and Editing for Translators*. 3rd eition. Abingdon: Routledge.

- **Newton, John** (1992). "Introduction and Overview." John Newton (ed.) (1992). *Computers in Translation: A Practical Appraisal*. London: Routledge, 1-14.

- **Núñez, Gabriel** (2013). "Translating for linguistic minorities in Northern Ireland: A look at translation policy in the judiciary, healthcare and local government." *Current Issues in Language Planning* 14(3-4), 474-489.

- **O'Brien, Sharon** (2006a). "Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output." *Across Languages and Cultures* 7(1), 1-21.

- — (2006b). *Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD Thesis. Dublin City University.

- — (2006c). Sharon O'Brien (2006) "Eye Tracking and Translation Memory Matches." *Perspectives –Studies in Translatology*, 14, 185-205.

- — (2007). "An Empirical Investigation of Temporal and Technical Post-editing Effort." *Translation and Interpreting Studies* 2(1), 83-136.

- — (2008). "Processing fuzzy matches in Translation Memory tools: an eye-tracking analysis." Susanne Göpferich, Arnt Lykke Jakobsen and Inger Mees (eds) (2008). *Looking at Eyes: Eye-tracking Studies of Reading and Translation Processing.* , Copenhagen: Samsfundslitteratur, 79-103.
- — (2011). "Towards predicting post-editing productivity." *Machine Translation* 25, 197-215.

- **O'Curran, Elaine** (2014). "Machine Translation and Post-Editing for User Generated Content: An LSP Perspective." Paper presented to the *11th Conference of the Association for Machine Translation in the Americas (AMTA)* (Vancouver, Canada October 22-26).

- **Ortiz-Boix, Carla & Matamala, Anna** (2016). "Post-editing wildlife documentary films: A new possible scenario?" *JoSTrans, The Journal of Specialised Translation*, 26, 187-210

- **Pavlović, Nataša & Jensen, Kristian** (2009). "Eye tracking translation directionality." Anthony Pym & Alexander Perekrestenko (eds) (2009). *Translation Research Projects 2.* Tarragona: Universitat Rovira i Virgili, 101-191.

- **Plitt, Mirko & Masselot, François** (2010). "A Productivity test of Statistical Machine Translation post-Editing in a typical localisation context." *The Prague Bulletin of Mathematical Linguistics* 93, 7-16.

- **Prys, Delyth, Prys, Gruffudd & Jones, Dewi** (2009). *Gwell Offer Technoleg Cyfieithu ar gyfer y Diwydiant Cyfieithu yng Nghymru: Arolwg Dadansoddol.* Bangor*:* Unedau Technoleg Iaith, Canolfan Bedwyr.

- **Pugh, Jeanette** (1992). "The story so far: The evaluation of machine translation in the world today." John Newton (ed.) (1992). *Computers in Translation: A Practical Appraisal.* London: Routledge, 14-33.

- **Ramos, Luciana** (2010). "Post-Editing Free Machine Translation: From a Language Vendor's Perspective." Paper presented at the *Ninth Conference of the Association for Machine Translation in the Americas* (Denver Colorado, October 31-November 4 2010).

- **Rothe-neves, Rui** (2003). "The influence of working memory features on some formal aspects of translation performance." Fabio Alves (ed.) (2003). *Triangulating Translation: Perspectives in Process Orientated Research.* Amsterdam/Philadelphia: John Benjamin's Publishing Company, 97-119.

- **Ruiz, Carlos. *et al.*** (2008). "Activation of lexical and syntactic target language properties in translation." *Acta Psychologica* 128(4), 490-500.

- **Screen, Benjamin** (2016). "What does Translation Memory do to translation? The effect of Translation Memory output on specific aspects of the translation process*." Translation and Interpreting* 8(1), 1-18.

- **Schilperoord, Joost** (1996). *It's About Time: Temporal Aspects of Cognitive Processes in Text Production*. Amsterdam: Rodopi

- **Sekino, Kyoko** (2015). "An Investigation of the Relevance-Theoretic Approach to Cognitive Effort in Translation and the Post-editing Process." *Translation and Interpreting* 7(1), 142-154.

- **Sharmin Selina, Spakov Oleg & Räihä, Kari-Jouko & Jakobsen, Arnt Lykke** (2008). "Where and for how long do translation students look at the screen while translating?" Susane Göpferich, Arnt Lykke Jakobsen & Inger Mees (eds) (2008). *Looking at Eyes: Eye-tracking Studies of Reading and Translation Processing.* Copenhagen: Samsfundslitteratur, 31-51.

- **Shreve, Gregory & Diamond, Bruce** (1997). "Cognitive Processes in Translation and Interpreting: Critical Issues." Joeseph Danks *et al*. (eds) (1997). *Cognitive Processes in Translation and Interpreting*. Thousand Oaks: Sage Publications, 233-252.

- **Silva, Roberto** (2014). "Integrating Post-editing MT in a Professional Translation Workflow." Sharon O'Brien *et al*. (eds) (2014). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing, 24-51.

- **Snover, Mathew & Dorr, Bonnie** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation". Paper presented to the *7th Conference of the Association for Machine Translation in the Americas (AMTA)* (Cambridge MA, August 8-12).

- **Somers, Harold** (2003). "Introduction." Harold Somers (ed.) (2003). *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamin's Publishing Company, 1-13.

- **TAUS** (2010) Post-editing in Practice. A TAUS Report. https://www.taus.net/think-tank/reports/postedit-reports/postediting-in-practice Accessed 7 December 2015

- **Techscribe** (2010). Evaluation of English-Welsh Machine Translation. https://www.docdroid.net/Dg02jzf/techscribe-eval.xls.html (consulted 12.03.2017).

- **Teixeira, Carlos** (2011). "Knowledge of provenance and its effects on translation performance in an integrated TM/MT environment". Paper presented to the *8th International NLPCS Workshop - Special theme: Human-Machine Interaction in Translatio*n. (Copenhagen, Denmark, October 20-21 2011).

- **Teixeira, Carlos** (2014). "Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories." Paper presented to the *Third Workshop on Post-Editing Technology and Practice (WPTP-3), within The eleventh biennial conference of the Association for Machine Translation in the Americas (AMTA-2014)*. (Vancouver, BC, Canada, October 26).

- **Turian, Joseph, Shen, Luke & Melamed, Dan** (2003). "Evaluation of Machine Translation and Its Evaluation". Paper presented to the *MT Summit IX* (New Orleans, September 2003).

- **Vandepitte, Sonia, Hartsuiker, Robert & Van Assche, Eva** (2015). "Process and Text Studies of a Translation Problem." Aline Ferreira & John Schwieter (eds) (2015). *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 127-145.

- **Vieira, Lucas** (2014). "Indices of cognitive effort in machine translation post-editing." *Machine Translation* 28, 187-216.

- **Watkins, Gareth** (2012). *Translation Tools and Technologies in the Welsh Language Context*. PhD Thesis. Swansea University.

- **Welsh Government** (2012). *A Living Language – A Language for Living, Welsh Language Strategy 2012-2017*. Cardiff: Welsh Government.

- — (2014). *A Living Language – A Language for Living, Moving Forward.* Cardiff: Welsh Government.

- **Zaretskaya, Anna, Corpas, Gloria & Seghiri, Miriam** (2015). "Integration of Machine Translation in CAT Tools: State of the Art, Evaluation and User Attitudes." *Skase Journal of Translation and Interpretation* 8(1), 76-89.

- **Zhechev, Ventsislav** (2014). "Analysing the Post-editing of Machine Translation at Autodesk." Sharon O'Brien *et al*. (eds) (2014). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing, 2-24.

Biography

Ben Screen is a PhD researcher at the School of Welsh, Cardiff University and an Associate Member of the Chartered Institute of Linguists. He has previously worked for a language service provider in Cardiff before returning to the School of Welsh in order to pursue research on effort, productivity and quality in the professional translation of Welsh when using translation tools.

Email: screenb@cardiff.ac.uk

Appendix 1: Source Text

Who can foster? If you have space in your home and time in your life you can help us make the big difference to a child's life. People do not need to be married to become a foster carer; they can be single, divorced, or living together. There are no upper age limits to becoming a foster carer but in Merthyr Tydfil we expect people to be sufficiently mature to work with children, some of whom can have complex needs. It is also expected that the foster carers are fit enough to provide for the child's needs. Our policy within Merthyr Tydfil is that we will not consider anyone who smokes to foster children under the age of 5 because of the associated health risks for children. How will the children I foster behave? If you expect the unexpected as you often might in caring for your own child, you won't go far wrong. It is not uncommon for children to feel withdrawn, insecure, or distressed when they arrive, and depending on circumstances this behaviour may be prolonged. Some children have been rejected or hurt by their parents and may be feeling angry, confused or anxious, so Foster Carers need to be prepared to allow such children to express themselves. This is best achieved in a safe and a secure environment, where the child's circumstances are considered and where clear boundaries can be set. Children who have been abused or harmed can display very disturbed behaviour and this can be daunting for anyone thinking of fostering. However, it is important to recognise that these are ordinary children who have suffered extraordinary circumstances and still need nurturing and their basic needs met, as with any other child. Expressing An Interest. If you would like more information or if you would like a member of our fostering team to contact you to discuss matters further please click here.

Appendix 2: Machine Translation for Post-editing

Pwy all faethu? Os oes gennych le yn eich cartref ac amser yn eich bywyd, gallwch ein helpu i wneud y gwahaniaeth mawr i fywyd plentyn. Nid oes angen i fod yn briod i fod yn ofalwr maeth bobl; gallant fod yn sengl, wedi ysgaru, neu'n byw gyda'i gilydd. Nid oes unrhyw derfynau oedran uchaf i fod yn ofalwr maeth, ond ym Merthyr Tudful yn disgwyl i bobl fod yn ddigon aeddfed i weithio gyda phlant, gall rhai ohonynt ag anghenion cymhleth. Disgwylir hefyd bod y gofalwyr maeth yn ddigon iach i ddarparu ar gyfer anghenion y plentyn. Mae ein polisi o fewn Merthyr Tudful yw na fyddwn yn ystyried unrhyw un sy'n ysmygu i feithrin plant o dan 5 oed oherwydd y risgiau iechyd sy'n gysylltiedig ar gyfer plant. Sut fydd y plant i faethu yn ymddwyn? Os ydych yn disgwyl yr annisgwyl wrth i chi aml efallai wrth ofalu am eich plentyn eich hun, ni fyddwch yn mynd yn bell o'i le. Nid yw'n anghyffredin i blant deimlo eu tynnu'n ôl, ansicr, neu'n ofidus pan fyddant yn cyrraedd, ac yn dibynnu ar amgylchiadau gall yr ymddygiad hwn fod yn hir. Mae rhai plant wedi cael eu gwrthod neu eu hanafu gan eu rhieni a all fod yn teimlo'n ddig, yn ddryslyd neu'n bryderus, felly mae angen i Ofalwyr Maeth fod yn barod i ganiatáu i blant o'r fath i fynegi eu hunain. Cyflawnir hyn orau mewn amgylchedd diogel ac amgylchedd diogel, lle mae

amgylchiadau'r plentyn yn cael eu hystyried a lle y gall ffiniau clir yn cael eu gosod. Gall plant sydd wedi cael eu cam-drin neu ei niweidio arddangos ymddygiad afreolaidd iawn a gall hyn fod yn frawychus i unrhyw un sy'n ystyried maethu. Fodd bynnag, mae'n bwysig cydnabod bod y rhain yn blant cyffredin sydd wedi dioddef amgylchiadau eithriadol ac yn dal i fod angen meithrin ac yn diwallu eu hanghenion sylfaenol, fel gydag unrhyw blentyn arall. Mynegi Diddordeb. Os hoffech gael mwy o wybodaeth neu os hoffech aelod o'n tîm maethu i gysylltu â chi i drafod materion ymhellach cliciwch yma.

---

[1] As noted by Zaretskaya, Pastor & Seghiri (2015), however, customised MT may in fact be out of reach for many freelancers or independent language service providers who may not have the time, resources or knowledge to be able to create or purchase such systems. Some smaller languages may also find it difficult to find sufficient data to feed the language and translation models in one particular domain. In these cases, Google Translate as well as other available generic systems may be a more practical option.

[2] To such an extent that the annual cost of translation between English and Welsh was estimated in 2007 to be worth £45,000,000 per annum (Prys, Prys & Jones 2007). Whilst this is a drop in the water compared to other European languages, it is still a significant sum and is likely to grow due to recent developments in Welsh language policy following the Welsh Language Measure (2011), which amongst other things made Welsh an official language in Wales.

[3] Previous published research has found a link between WMC and professional translation. Despite finding no statistically significant regression model between features of working memory and the translation processes of novices, Rothe-Neves (2003) did in fact find such a link between professional translation and working memory. Vieira (2014) also found a similar result for post-editing.

[4] This tendency to self-revise before completing a final draft (one of the two main forms of translation revision according to Mossop (2013), has been found to be a feature of translation in two process studies. Alves & Vale (2011) and Dragsted & Carl (2013) both found different revision strategies displayed by participants, one of which was self-revision after forming a first draft. It is interesting that this has also been found in this study, as well as being found to be a feature of post-editing.